**Instructions**

You have 4.5 hours to complete the exam.

You may use R, Stata, Python, SPSS, SAS, Excel or any program(s) of your choice to assist you, but please paste your answers (including model outputs) into Microsoft Word.

You may also look online, use notes, textbooks, software instructions, any software of your choice. Please indicate at the end of your exam what materials you used.

You may **NOT** ask anyone else for help with this exam. You may **NOT** include any content that is not your own. Also, do **NOT** include any content from previous work, classes, etc. All answers must be original.

Suggested timing for each question is just a guide – you may spend more time with some questions and less time with others, provided that you return the completed exam within four hours after the exam is emailed to you. The suggested time also indicates the relative value of each question.

Question 1 – 30 minutes

Question 2 – 40 minutes

Question 3 – 15 minutes

Question 4° – 45 minutes

Question 5\* – 85 minutes

Question 6† – 55 minutes

°Supplemental data for question 4 is provided in “poll\_responses.csv”

\*Question 5 is based on data provided in “model\_data.csv”

†Question 6 is based on data provided in “work\_program\_data.csv”

**Question 1**

You’re working for a statewide Democratic campaign. The campaign has access to a database with a record for each individual registered voter in the state – including their name, address, party registration, past vote history, demographic information and more.

This information has been combined with a recent telephone poll of 5K random constituents – where each person was asked what candidate they planned on supporting: Democrat Herman Madison or Republican Martha Whistler. There are no other candidates.

Using this combined dataset, one of your fellow analysts has built a *logistic regression* model that predicts the probability an individual voter will support the Democrat.

Below are coefficients from this model. The definitions of the variables are below.

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **coefficient** | **standard error** | **z score** |
| *Democrat* | 1.45 | 0.09 | 15.81 |
| *Republican* | -2.11 | 0.1 | -21.95 |
| *Ln\_Income* | -0.109 | 0.041 | -2.63 |
| *Age* | -0.013 | 0.0096 | -1.4 |
| *Age\_Sq* | 0.0001 | 0.00009 | 1.52 |
| *Census\_College* | 1.77 | 0.33 | 5.37 |
| *AfAm* | 2.07 | 0.399 | 5.18 |
| *AfAm\_Democrat* | -0.872 | 0.437 | -2.01 |
| *Constant* | 1.2 | 0.484 | 2.48 |

* Democrat—Coded as 1 if the voter is a registered Democrat, 0 if he/she is not
* Republican—Coded as 1 if the voter is a registered Republican, 0 if he/she is not
* Ln\_Income—The natural logarithm of the voter’s income (in dollars)
* Age—The voter’s age (in years)
* Age\_Sq—The voter's age (in years) squared
* Census\_College—The percentage of residents in the voter’s neighborhood who have a college degree (scaled from 0 to 100)
* AfAm—Coded as 1 if the voter is African American, 0 if he/she is not
* AfAm\_Democrat—Coded as 1 if the voter is both African American and a registered Democrat, 0 if he/she is not
* Constant—The constant term

**Part A**  
Consider 4 voters, Adam, Bob, Chris and David. Adam and Chris share identical characteristics except for their incomes.  Bob and David also share identical characteristics (with each other, not necessarily Adam and Chris), except for their incomes.

|  |  |  |
| --- | --- | --- |
| **Name** | **Income** | **Modeled Support** |
| Adam | $50,000 | 50% |
| Bob | $200,000 | 50% |
| Chris | $40,000 | ? |
| David | $190,000 | ? |

Based on the coefficients above, who would you think has a higher probability of supporting Herman Madison?

* Chris
* David
* They have the same probability
* Cannot tell based on the information provided

What is your reasoning?  (you need not calculate an exact probability to answer this question. Just explain your reasoning in general terms.)

**Part B**  
The coefficient for AfAm\_Democrat is negative. How do you interpret this? Does this mean that African-American registered Democrats support Herman Madison at lower rates than African-American independents? What about relative to registered Democrats of other ethnicities?

**Part C**  
How do we interpret the difference in support between voters of different ages? How do the variables in the model estimate such support?

**Part D**  
Are there any variables in this model that you would choose to drop? Why or why not? Would you need more information in order to make this decision?

**Question 2**

Below is the family tree for a small parade of elephants, one of which (“C”) exhibits albinism caused by a mutation in the gene for skin pigmentation. Each elephant carries two copies of this gene and in order to be an albino, both copies must have the mutation. When the elephants mate, each parent passes one of their copies onto the child with equal probability (regardless of mutation).

Elephants A and B each carry only 1 copy of the mutation (and thus don’t exhibit albinism) and elephant C carries 2 copies (and is affected), as shown in the figure. Elephants C, D, and E are all descendants of A & B, and Elephant F is a descendant of D & E.



*Please show your work or reasoning for each of the following questions:*

**Part A.** Given that elephant D is not an albino, what is the probability it has exactly one copy of the mutation?

**Part B.** Given that elephant E is not an albino, what is the probability it has exactly one copy of the mutation?

**Part C.** What is the probability that elephant F, when born, will be an albino?

**Part D.** What is the probability that elephant F, when born, will carry exactly one copy of the mutation?

**Part E.** Suppose that elephant F is born and is not an albino. Given this, what is the probability that elephant F carries exactly one copy of the mutation?

**Part F.** Suppose that elephant F is born and is not an albino. Given this, what is the probability that elephant D carries exactly one copy of the mutation?

**Question 3**  
  
You’re working in the analytics department of Democrat Zoey Bartlet’s campaign for Congress in New Hampshire. It’s September and you’ve just built a statistical model to predict each individual registered voter’s probability of supporting Bartlet (based on voter and survey data).   
  
In your model, each voter is assigned a support score ranging from 0 to 100. You’ve shared these scores with everyone on the campaign.  
  
One day, you receive an email from the campaign manager asking the following question:

“According to your model, my friend Ruby has a support score of 50 (out of 100), but I’ve known Ruby for 20 years, and she always votes for Democratic candidates. There must be something wrong with your model if it says she only has a 50% chance of voting for Zoey.”

How would you respond to the campaign manager? Assume that she is a smart, educated person with extensive political experience and little or no background in statistics.

**Question 4**  
  
You have recently started in the market research division of a box company in Tustin, CA. As one of your first tasks, Randy, the head of market research, asks you to analyze the following data from a recent poll of 5,000 Californians:

*Question: If you were to purchase boxes in the next few months, would you prefer corrugated or uncorrugated cardboard for your boxes?*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **gender** | **age\_bucket** | **urbanicity** | **California Population** | **Respondents** | **Prefer Corrugated** |
| male | 18\_to\_35 | urban | 4,815,108 | 252 | 164 |
| female | 18\_to\_35 | urban | 4,151,623 | 260 | 145 |
| male | 18\_to\_35 | rural | 2,342,416 | 234 | 156 |
| female | 18\_to\_35 | rural | 1,854,720 | 228 | 123 |
| male | 36\_to\_64 | urban | 6,676,992 | 678 | 451 |
| female | 36\_to\_64 | urban | 6,259,680 | 714 | 373 |
| male | 36\_to\_64 | rural | 3,338,496 | 684 | 434 |
| female | 36\_to\_64 | rural | 2,543,616 | 568 | 293 |
| male | 65\_plus | urban | 1,516,234 | 354 | 225 |
| female | 65\_plus | urban | 1,608,023 | 428 | 224 |
| male | 65\_plus | rural | 741,888 | 260 | 168 |
| female | 65\_plus | rural | 927,360 | 341 | 187 |

For your convenience, the individual-level responses from the poll are provided in “poll\_responses.csv” which you should have received along with this exam (yes responses are coded as 1 and no responses as 0 in the file). The questions below can be answered with or without this additional data.

**Part A.** What is your best estimate for the percentage of Californians who prefer corrugated cardboard for their boxes? What is the 95% confidence interval for this estimate? Please show your work.  
  
**Part B.** Are men and women significantly different in their likelihood to prefer corrugated cardboard?  
  
**Part C.** Your company is currently planning for 54% of the boxes it makes next quarter to use corrugated cardboard. How (if at all) would you recommend adjusting production in the upcoming quarter based on the results of your analysis? Explain your reasoning.

**Question 5**

It’s 28 days until election day and you’re working for a Democratic candidate for Senate. Your team has been asked to build a statistical model to help identify likely supporters for the campaign’s “Get Out the Vote” operation.

The campaign’s data team has assembled the dataset “model\_data.csv” and your task is to use it to build this support score model. There are two steps:

**PART A:** Model building – you will build a model using some or all of the attached data (*consider part B before starting part A*).

**PART B:** Validation – you will validate this model using some or all of the attached data.

**PART A**

For your convenience, we have put the data into a CSV file (attached). You may import or copy and paste this data into any statistics package of your choice (Stata, R, SAS, SPSS) to build your model.

Your job is to produce a simple model that predicts the probability of identifying for the Democratic candidate based on the attached data. We have also included a **data dictionary on the last page of this exam** that defines each variable for your reference.

Feel free to use not only the variables included in the attached data set, but also other variables built upon these (such as interactions or transformations).

* The data may have some missing values. Please keep this in mind, and explain how you will deal with this missing data and missing data in general.
* Please tell us what kind of models and algorithms you would consider and explain your choice of the model you decided to build.
* Once you have selected a single model type (regression, decision tree, support vector machine, etc.), please build at least two different variations of that model. For example, you may want to vary which variable(s) are included, or you may want to try a variable transformation or interaction. Please copy and paste the results of each variation into your MS Word document. Discuss why your final model is superior to other models you tried. For your final model, please explain what variables are most important and how the results should be interpreted.
* Please describe one or more graphics you could generate to use as a diagnostic tool to evaluate the quality of the model or as a visual tool to demonstrate the effectiveness of the model. (You may create one or more of these graphics if you have extra time, but it is not necessary.)
* If you had more time, what else would you do? What other variables would you ask for or want to construct? What other model specifications would you explore and why (briefly)?
* Additionally, using your final model please create a column on the CSV that gives a probability that each voter will support the Democratic candidate (**return the resulting dataset with your exam**). Please note that we would like scores for all voters, including those for whom the dependent variable (support\_democrat) is missing.

**PART B**

Use some or all of the attached data to validate your model.

* How well does your model validate? Why do you say that?
* Now suggest another way you could validate your model using external data rather than the attached data. What additional value would this validation provide?

**Question 6**  
  
You should have received a file named “work\_program\_data.csv” along with this exam. This file contains data for participants in a subsidized work program that was run during the mid-1970’s as well as a sample of individuals who did not participate in the program. The file contains the following variables:

participant =1 for program participants, 0 otherwise  
age age in years  
educ years of education  
afam =1 if African American, 0 otherwise  
hisp =1 if Hispanic, 0 otherwise  
white =1 if white, 0 otherwise  
married =1 if married, 0 otherwise  
income\_1974 real (inflation adjusted) income for 1974 (pre-program)  
income\_1975 real (inflation adjusted) income for 1975 (pre-program)  
income\_1978 real (inflation adjusted) income for 1978 (post-program)  
unemployed\_1974 =1 if unemployed in 1974 (pre-program), 0 otherwise  
unemployed\_1975 =1 if unemployed in 1975 (pre-program), 0 otherwise  
unemployed\_1978 =1 if unemployed in 1978 (post-program), 0 otherwise

The program was open to the population and participation was based solely on individuals’ choice to do so.

**Part A**  
Provide an estimate of the effect of participating in this program on earnings in 1978. You may import or copy the data from the CSV into any statistics or data analysis program of your choice, but please include a copy of any code you wrote to analyze the data as an attachment to your exam.  
  
**Part B**  
Explain the methodology you chose to estimate this program’s effect. What are its strengths and weaknesses?

**Affirmation**

1. Please indicate what tools (Websites, textbooks, class notes, software, etc.) you used during the course of this exam and how you used them.
2. Please type your name, location, time and date below to affirm that you neither sought nor received assistance from anyone else during this exam and that all answers are original and entirely your own work and do not include any content created prior to this four hour exam.

“I hereby affirm that I neither sought nor received assistance from anyone else during this exam and that all answers are original and entirely my own work and do not include any content created prior to this four hour exam.”

Name \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Location \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Time \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Date \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Data Dictionary for Question 5**

|  |  |
| --- | --- |
| Variable Name | Definition |
| id | Unique identifier for each voter |
| support\_democrat | 1 if the voter intends to support the Democrat, 0 if s/he intends to support the Republican (Dependent Variable) |
| cong\_district\_region | The region of the congressional district in which the voter resides |
| occupation\_code | Modeled occupation: 1=Managerial. 2=Professional. 3=Service. 4=Clerical. 5=Technical. 6=Agriculture. 7=Industrial. 8=Technology. 9=Retail. 0=Other. |
| gender\_female | 1 for females, 0 for males |
| age | Voter's age in years |
| party | A voter's party registration. 1=Democrat. 2=Republican. 3=Green Party. 4=Libertarian Party. 5=Independent/Other |
| census\_urbanpcnt | Percentage of neighborhood that is urban |
| census\_suburbanpcnt | Percentage of neighborhood that is suburban |
| census\_ruralpcnt | Percentage of neighborhood that is rural |
| census\_collegepcnt | Percentage of neighborhood that is college graduates |
| census\_unemprate | Unemployment rate in the neighborhood |
| census\_medianincome | Median income in the neighborhood |
| density\_sq\_km | Population density of the neighborhood |
| on\_email\_list | Signed up for the campaign’s email list |
| avg\_dem\_performance | Percentage of a precinct that votes for candidate on average |
| pet\_owner | Owns a pet |
| golf | Likes golf |
| hunting | Likes hunting |
| random | Random integer between 1 and 10 |
| likes\_cheese | Outside model predicts that this voter enjoys the taste of cheese |
| protestant | Religion (Protestant) |
| catholic | Religion (Catholic) |
| jewish | Religion (Jewish) |
| afam | Ethnicity (African-American) |
| latino | Ethnicity (Latino) |
| id\_d | When surveyed last cycle about their preference in the Governor's race, they supported the Democratic candidate |
| id\_r | When surveyed last cycle about their preference in the Governor's race, they supported the Republican candidate |
| score\_demo | Support score created using the demographic and neighborhood level variables in this file. This score was modeled on the same dependent variable available to you. |